STATISTICAL INFERENCE
Lecture 1
Skeleton notes

1 Role of theory of inference

Objective is to provide concepts and methods helpful for science, technology, public affairs, etc. Very wide variety of problems require variety of approaches. Ultimate criterion is relevance.

Idealized scheme:

- research question or questions
- study design
- data collection
- preliminary analysis
- more formalized probabilistic analysis
- conclusions and interpretation and usually
- more questions

Formal theory of inference needed to underpin and systematize methods and to provide base for tackling new problems. In data mining, and to some extent more generally, formalizing the right question is one of the objectives.

2 Probabilistic formulation

Assume observations on response (outcome) variables and explanatory variables. Typically treat the former as observed values of a random vector Y having a distribution depending on the explanatory variables regarded as fixed, the distribution specified by a model $f_Y(y;\theta)$. giving p.d.f. of Y as a function of known x, omitted from notation, and unknown parameter vector θ.

Usually θ is partitioned (ψ, λ) into parameter of interest ψ and nuisance parameter λ. Model is an idealized model of variation in the physical, biological, $\ldots$, world and probabilities represent limiting frequencies under (often hypothetical) repetition.

Model choice is of key importance. It translates a subject-matter question into a statistical one. Sometimes model represents data-generating process, in others it is largely empirically descriptive. Parameters aim to capture features of the system under study separated from features specific to the particular data. Choice of parameters of interest crucial.

There are now a number of possible objectives:

- various possibilities studied on the basis that the model is sound
- model criticism

Specific objectives include the following

- what can be concluded about the value of ψ?

- reach a decision among possibilities whose merit depends on ψ
- predict the value of a new observation from the same or related systems
- is there need for model change? Model criticism.

Strategical aspects of how to use statistical methods not considered here.

3 Broad approaches

There are two main formal approaches to these issues

- frequentist in which probability is constrained to mean a (usually hypothetical) frequency
- inverse probability (Bayesian) in which often the notion of probability is extended to cover assessment of (any) uncertain event or proposition

Both approaches have a number of variants. In some but by no means all situations the numerical answers from the two approaches are nearly or even exactly the same, although the meanings are even then subtly different.

4 Examples

In the simplest example $Y_1, \ldots, Y_n$ are iid with a normal distribution of unknown mean μ and known variance σ_0^2. In the general notation μ is the parameter of interest ψ. Had the variance been unknown it would have been a nuisance parameter. Of course the definition of the parameter of interest depends totally on the research question. With two unknown parameters the parameter of interest might, for example, have been μ/σ, although most commonly it is μ.

In the lecture the following example will be used to illustrate general issues: the random variables $Y_1, \ldots, Y_n$ are iid with the exponential distribution of rate parameter ρ, i.e. with mean $\mu = 1/\rho$,

There are now a variety of problems corresponding to different questions and to different approaches.

5 Exponential mean

5.1 Initial analysis

First step: find likelihood

Exponential family

Sufficient statistic, $s = \Sigma y_l$

Key to importance of sufficiency

The parting of the ways!

- frequentist: what is the probability distribution of $S = \Sigma Y_l$ for fixed value of the known constant ρ?
- Inverse probability (Bayesian) approach. Value of ρ is unknown and therefore has a probability distribution with and without the data. That is, ρ is the value of a random variable P.

In general a *pivot* is a function, $p(y, \psi)$ of the data y and parameter of interest ψ which has a fixed distribution and which is monotonic in ψ for every fixed y. In frequentist theory we consider the distribution of $p(Y, \psi)$ for each fixed θ, whereas in Bayesian theory we consider the distribution of $p(y, \Psi)$ for each fixed y. Common form of pivot is that of an estimate minus the parameter value divided by a standard error.

6 Frequentist approach

Analyses and measures of uncertainty calibrated by performance under hypothetical repetition

- simple significance test
  - modelled on testing a deterministic hypothesis
- test, Neyman-Pearson style
- confidence intervals
- prediction

7 Simple significance test

Deterministic hypothesis tested

- find interesting observable consequence of hypothesis
- collect observation
- - consistency with hypothesis
  - inconsistent

  Statistical null hypothesis tested

  - find interesting aspect, t, of data whose distribution under hypothesis is known
  - arrange that large values of t correspond to departures as before
  - collect data
  - calculate t

– find

$$p = P(T \geq t; H_0).$$

– application

– interpretation; if we were to regard the current data as just decisive evidence against H_0 then in a long run of applications in which hypothesis true we would be wrong in a proportion p of times.

8 Test of hypothesis; Neyman-Pearson style

– require formulation of probability model for H_0 and one or more alternatives H_A

– for given α find set of values with probability at most α under H_0 and in some sense maximum probability under the alternatives. Equivalent to choice of test statistic

– reject or accept H_0 according to whether data fall or do not fall in region in question

– in theoretical formulations α is a pre-chosen constant but in practice implementation is often closer to simple sig. test

9 Confidence intervals or limits

– Direct argument from pivot

– Set of parameter values consistent with data up to specified significance level

10 Bayesian approach

All calculations by laws of probability.

Leads to posterior density of P, the random variable corresponding to ρ.

But what do the answers mean? Tests: two Bayesian versions

- Atom of probability at H_0
- Prior over alternatives must be formulated

Sometimes better interpreted in terms of the question: does the apparent effect have the wrong sign?

EXERCISE

Suppose that s^2 is the residual mean square with d_{res} degrees of freedom in a normal theory linear model and σ^2 is the true variance. Suppose that it is decided to base inference about σ^2, whether Bayesian or frequentist, solely on s^2. You may assume that the random variable S^2 is such that $d_{\mathrm{res}}S^2/\sigma^2$ is distributed as chi-squared with d_{res} degrees of freedom.
(i) What is the 95 per cent upper confidence limit for σ? (ii) For large d the chi-squared distribution with d degrees of freedom is approximately normal with mean d and variance $2d$. How large would d_{res} have to be for the 95 percent upper limit to be $1.2s_{\mathrm{res}}$? (iii) What is the conjugate prior in a Bayesian analysis? When, if ever, would posterior and confidence limits agree?

STATISTICAL INFERENCE
Lecture 2
Skeleton notes

1 Use of a minimal sufficient statistic: some principles

Here 'sufficient statistic' will always mean *minimal* sufficient statistic.

Notation:

- random vector Y
- parameter (usually vector) θ
- sometimes $\theta = (\psi, \lambda)$, with ψ of interest and λ nuisance
- symbol f used for pdf, pmf — conditional or marginal as indicated by context (and sometimes explicitly by subscripts).

1.1 Inference on θ

Sufficient statistic S:

$$f(y; \theta) = f_S(s(y); \theta) f_{Y|S}(y|s)$$

where the second factor does not involve θ.

Implications:

- inference for θ based on $f_S(s;\theta)$
- $f_{Y|S}(y|s)$ eliminates θ, and provides a basis for model checking.

Idea here is that S is a substantial reduction of Y.

(At the other extreme, if the minimal sufficient statistic is $S = Y$, the second factor above is degenerate and this route to model-checking is not available.)

1.2 Inference on ψ (free of λ)

Often $\theta = (\psi, \lambda)$, where ψ is the parameter (scalar or vector) of interest, and λ represents one or more nuisance parameters.

Ideal situation: there exists statistic S_λ — a function of the minimal sufficient statistic S — such that, for every fixed value of ψ, S_λ is sufficient for λ. For then we can write

$$f(y;\psi,\lambda) = f_{Y|S_\lambda}(y|s_\lambda;\psi)f_{S_\lambda}(s_\lambda;\psi,\lambda),$$

and inference on ψ can be based on the first factor above.

This kind of factorization is not always possible. But:

- exponential families — exact;
- more generally — approximations.

1.3 Inference on model adequacy (free of θ)

How well does the assumed model $f_Y(y;\theta)$ fit the data?

Now θ is the 'nuisance' quantity to be eliminated.

Suppose that statistic T is designed to measure lack of fit. Ideally, T has a distribution that does not involve θ: a significant value of T relative to that distribution then represents evidence against the model (i.e., against the *family* of distributions $f_Y(y;\theta)$).

Condition on the minimal sufficient statistic for θ: refer T to its conditional distribution $f_{T|S}(t|s)$, which does not depend on θ.

2 Exponential families

Introduced here as the cleanest/simplest class of models in which to explore and exemplify the above principles.

2.1 Introduction: some special types of model

Many (complicated) statistical models used in practice are built upon one or more of these three types of family:

- transformation family;
- mixture family;
- exponential family.

Transformation families and exponential families are excellent models for the purpose of studying general principles. (Mixture families tend to be messier, inferentially speaking.)

Our main focus in the rest of this lecture will be on exponential families. The other two types will be introduced briefly for completeness.

2.1.1 Transformation families

Prime examples of a transformation model are

- *location model* $f(y;\theta) = g(y-\theta)$
- *scale model* $f(y;\theta) = \theta^{-1} g(y/\theta)$
- *location-scale model* $f(y;\mu,\tau) = \tau^{-1} g\{(y-\mu)/\tau\}$

where in each case $g(.)$ is a fixed function (not depending on θ).

Each such model is characterized by a specified group of transformations.

2.1.2 Mixture families

Simplest case: 2-component mixture

$$f(y;\theta) = (1-\theta) f(y;0) + \theta f(y;1) \qquad (0 \le \theta \le 1),$$

where $f(y;0)$ and $f(y;1)$ are the specified 'component' distributions.

More generally: any number of components (possibly infinite), with θ indexing a suitable 'mixing' distribution.

Summation of components makes life easy in some respects (normalization is automatic), but much harder in other ways (no factorization of the likelihood).

2.1.3 Exponential families

When the parameter is the canonical parameter of an EF, we will call it ϕ instead of θ (merely to remind ourselves).

An EF interpolates between (and extrapolates beyond) component distributions on the scale of $\log f$ (cf. mixtures; interpolation on the scale of f itself). For example, a one-parameter EF constructed from two known components is $f(y;\theta)$ such that

$$\begin{aligned} \log f(y;\phi) &= (1-\phi)\log f(y;0) + \phi \log f(y;1) - k(\phi) \\ &= \phi \log \frac{f(y;1)}{f(y;0)} + \log f(y;0) - k(\phi), \end{aligned}$$

where the $k(\phi)$ is needed in order to normalize the distribution. This is an instance of the general form for an EF (see the preliminary material)

$$f(y;\phi) = m(y)\exp[s^T(y)\phi - k(\phi)].$$

Some EFs are also transformation models [but not many! — indeed, it can be shown that among univariate models there are just *two* families in both classes, namely $N(\mu, \sigma^2)$ (a location-scale family) and the Gamma family with known 'shape' parameter α (a scale family)].

2.2 Canonical parameters, sufficient statistic

Consider a d-dimensional full EF, with canonical parameter vector $\phi = (\phi_1, \ldots, \phi_d)$, and sufficient statistic $S = (S_1, \ldots, S_d)$.

Clearly (from the definition of EF) the components of ϕ and of S are in one-one correspondence.

Suppose now that $\phi = (\psi, \lambda)$, and that the corresponding partition of S is $S = (S_\psi, S_\lambda)$.

It is then immediate that, for each fixed value of ψ, S_λ is sufficient for λ. This is the 'ideal situation' mentioned in 1.2 above.

More specifically:

1. the distribution of S is a full EF with canonical parameter vector ϕ;

2. the conditional distribution of S_ψ, given that $S_\lambda = s_\lambda$, is a full EF with canonical parameter vector ψ.

2.3 Conditional inference on parameter of interest

The key property, of the two just stated, is the second one: the conditional distribution of S_ψ given S_λ is free of λ. This allows 'exact' testing of a hypothesis of the form $\psi = \psi_0$, since the null distribution of any test statistic is (in principle) known — it does not involve the unspecified λ.

Tests $\rightarrow$ confidence sets.

Note that the canonical parameter vector ϕ can be linearly transformed to $\phi' = L\phi$, say, with L a fixed, invertible $d \times d$ matrix, without disturbing the EF property:

$$s^T\phi = [(L^{-1})^T s]^T (L\phi),$$

so the sufficient statistic after such a re-parameterization is $(L^{-1})^T S = S'$, say. This allows the parameter of interest ψ to be specified as any linear combination, or vector of linear combinations, of $\phi_1, \ldots, \phi_d$.

2.3.1 Example: 2 by 2 table of counts

Counts R_{ij} in cells of a table indexed by two binary variables:

$$\begin{array}{cc|c} R_{00} & R_{01} & R_{0+} \\ R_{10} & R_{11} & R_{1+} \\ \hline R_{+0} & R_{+1} & R_{++} = n \end{array}$$

Several possible sampling mechanisms for this:

- Individuals counted into the four cells as result of random events over a fixed time-period. Model: $R_{ij} \sim \text{Poisson}(\mu_{ij})$ independently. [No totals fixed in the model.]

- *Fixed* number n of individuals counted into the rour cells. Model: $(R_{00}, R_{01}, R_{10}, R_{11}) \sim \text{Multinomial}(n; \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$. [Grand total, n, fixed in the model]

- Row variable is treatment (present/absent), column variable is binary response. Numbers treated and untreated are fixed ($R_{0+} = n_0$, $R_{1+} = n_1$, say). Model: $R_{i0} \sim \text{Binomial}(n_i; \pi_i)$ $(i = 0, 1)$. [Row totals fixed in the model]

In each case the model is a full EF. Take the (canonical) parameter of interest to be

$$\psi = \log \frac{\mu_{11}\mu_{00}}{\mu_{10}\mu_{01}},$$

where $\mu_{ij} = E(R_{ij})$. In the pair-of-binomials model this is the log odds ratio.

In each case the relevant conditional distribution for inference on ψ turns out to be the same. It can be expressed as the distribution of R_{11}, say, conditional upon the observed values of all four marginal totals $M = \{R_{0+}, R_{1+}, R_{+0}, R_{+1}\}$:

$$\text{pr}(R_{11} = r_{11}|M) = \frac{\binom{r_{0+}}{r_{01}}\binom{r_{1+}}{r_{11}}\exp(r_{11}\psi)}{\sum \binom{r_{0+}}{r_{+1}-w}\binom{r_{1+}}{w}\exp(w\psi)}$$

— a *generalized hypergeometric* distribution.

When $\psi = 0$, this reduces to the ordinary hypergeometric distribution, and the test of $\psi = 0$ based on that distribution is known as *Fisher's exact test*.

The practical outcome (condition on all four marginal totals for inference on ψ) is thus the same for all 3 sampling mechanisms. But there are two distinct sources of conditioning at work:

Conditioning by model formulation: the multinomial model conditions on n; the pair-of-binomials model conditions on $r_{0+} = n_0, r_{1+} = n_1$.

'Technical' conditioning (to eliminate nuisance parameters) applies in all 3 models; the numbers of nuisance parameters eliminated are 3, 2 and 1 respectively.

2.3.2 Example: Several 2 by 2 tables

(The Mantel-Haenszel procedure)

Extend the previous example: m independent 2×2 tables, with assumed common log odds ratio ψ.

Pair-of-binomials model for each table: canonical parameters (log odds) for table k are

$$\phi_{k0} = \alpha_k, \qquad \phi_{k1} = \alpha_k + \psi.$$

Parameters $\alpha_1, \ldots, \alpha_m$ are nuisance. Eliminate by (technical) conditioning on all of the individual column totals, as well as conditioning (as part of the model formulation) on all the row totals.

Resulting conditional distribution is the distribution of $S_\psi = \sum R_{k.11}$ conditional upon all row and column totals — the convolution of m generalized hypergeometric distributions.

In practice (justified by asymptotic arguments), the 'exact' conditional distribution for testing $\psi = 0$ — the convolution of m hypergeometrics — is usually approximated by the normal with matching mean and variance.

2.3.3 Example: binary matched pairs

Extreme case of previous example: row totals $r_{k.0+}, r_{k.1+}$ are all 1.

Each table is a pair of independent *binary* observations (e.g., binary response before and after treatment).

Conditional upon column totals: only 'mixed' pairs k, with $r_{k.+0} = r_{k.+1} = 1$, carry any information at all.

Conditional distribution for inference on ψ is binomial. (see exercises)

This is an example where conditional inference is a *big* improvement on standard approximations based on the unconditional likelihood: e.g., the unconditional MLE $\hat{\psi}$ is inconsistent as $m \to \infty$, its limit in probability being 2ψ rather than ψ.

2.4 Conditional test of model adequacy

The principle: refer any proposed lack-of-fit statistic to its distribution conditional upon the minimal sufficient statistic for the model parameter(s).

We mention here just a couple of fairly simple examples, to illustrate the principle in action.

2.4.1 Example: Fit of Poisson model for counts

(Fisher, 1950)

Testing fit of a Poisson model.

Conditional distribution of lack-of-fit statistic given MLE (which is minimal sufficient since the model is a full EF).

Calculation quite complicated but 'do-able' in this simple example.

2.4.2 Example: Fit of a binary logistic regression model

A standard lack-of-fit statistic in generalized linear models is the *deviance*, which is twice the log likelihood difference between the fitted model and a 'saturated' model.

In the case of independent binary responses y_i the deviance statistic for a logistic regression with maximum-likelihood fitted probabilities $\hat{\pi}_i$ is

$$\begin{aligned} D &= 2\sum\left\{y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right\} \\ &= 2\sum\left\{y_i \log y_i + (1-y_i)\log(1-y_i) - y_i \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) - \log(1-\hat{\pi}_i)\right\} \end{aligned}$$

Since y is 0 or 1, the first two terms are both zero. Since the fitted log odds is $\log\{\hat{\pi}_i/(1-\hat{\pi}_i)\} = x_i^T\hat{\beta}$, the deviance can be written as

$$D = -2\hat{\beta}^T X^T Y - 2\sum\log(1-\hat{\pi}_i)$$

$$= \quad -2\hat{\beta}^T X^T \hat{\pi} - 2\sum \log(1 - \hat{\pi}_i),$$

since the MLE solves $X^T Y = X^T \hat{\pi}$.

Hence D in this (binary-response) case is a function of $\hat{\beta}$, which is equivalent to the minimal sufficient statistic.

The required conditional distribution of D is thus *degenerate*. The deviance statistic carries *no information at all* regarding lack of fit of the model.

The same applies, not much less severely, to other general-purpose lack of fit statistics such as the 'Pearson chi-squared' statistic $X^2 = \sum (y_i - \hat{\pi})^2 / \{\hat{\pi}_i(1 - \hat{\pi}_i)\}$.

This (i.e., the case of binary response) is an extreme situation. In logistic regressions where the binary responses are *grouped*, the lack-of-fit statistics usually have non-degenerate distributions; but when the groups are small it will be important to use (at least an approximation to) the conditional distribution given $\hat{\beta}$, to avoid a potentially misleading result.

Exercise

For the binary matched pairs model, derive the conditional binomial distribution for inference on the common log odds ratio ψ. Discuss whether it is reasonable to discard all the data from 'non-mixed' pairs.

STATISTICAL INFERENCE
Lecture 3
Skeleton notes

1 Brief assessment

In the model, probability is an idealized representation of an aspect of the natural world and represents a frequency.
Two approaches:

- Frequentist theory uses frequentist view of probability indirectly to calibrate significance tests, confidence intervals, etc

- Bayesian theory uses probability directly by typically using a different or more general notion of probability.

2 Frequentist theory

- covers a wide range of kinds of formulation

- provides a clear link with assumed data generating process

- very suitable for assessing methods of design and analysis in advance of data

- accommodates model criticism

- some notion of at least approximately correct calibration seems essential

but

- derivation of procedures may involve approximations, typically those of asymptotic theory
- nature of asymptotic theory
- there may be problems in specifying the set of hypothetical repetitions involved in calculating error-rates appropriate for the typically unique set of data under analysis
- use of probability to assess uncertainty is indirect

3 Bayesian approaches

- all calculations are applications of the laws of probability: find the conditional distribution of the unknown of interest given what is known and assumed
- if unknown is not determined by stochastic process, probability has to be a measure of uncertainty not directly a frequency

Central issues

- What does such a probability mean, especially for the prior?
- How do we determine numerical values for the prior?

- Bayesian frequentist theory (empirical Bayes)
  - role of hyperparameter
- impersonal (objective) degree of belief
- personalistic degree of belief

Objectives

- may be valuable way of inserting new evidence, for example expert opinion
- in other contexts interest may lie in a neutral or reference prior so that contribution of data is emphasized

but

- flat priors sometimes, but by no means always, in some sense represent initial ignorance or indifference
- most foundational work on Bayesian theory rejects the notion that a prior can represent an initial state of ignorance
- nominally a closed world
- issues of temporal coherency
- merges different sources of information without examining mutual consistency
- if meaning of prior is unclear so is that of posterior.

4 Some issues in frequentist theory

Central issue of principle (although not of practice) is how to ensure frequentist probability, an aggregate property, relevant to a unique situation.

Role of conditioning

5 Probability as a degree of belief

- impersonal (objective) degree of belief
- personalistic degree of belief
  - assessed in principle by Your betting behaviour
  - tied to personal decision making
  - for public discussion prior needs to be evidence-based
  - temporal coherency
  - mutual consistency of data and prior
  - escape from too narrow a world
  - model criticism

Six views of Bayesian approaches

- empirical Bayes
- objective degree of belief or standardized reference priot
- personalistic degree of belief
- technique for incorporating additional information

- personal decision making
- technical device for producing good confidence intervals

EXERCISE

The random variables $Y_1, \ldots, Y_n$ are independently normally distributed with unit variance and unknown means and n is large. It is possible that all the means are zero; alternatively a smallish number of the means are positive. How would you proceed from a Bayesian and from a frequentist perspective?

OR

The observed random variable Y is normally distributed with mean μ and unit variance. The prior distribution of μ assigns equal probability 1/2 to the values ± 10. We observe $y = 1$. What would be concluded about μ?

STATISTICAL INFERENCE
Lecture 4
skeleton notes

Scalar parameter

Score function:

$$U = \frac{\partial l(\theta; Y)}{\partial \theta}$$

— a random function of θ.

The score has mean zero at the true value of θ (subject to regularity condition).

Regularity: can validly differentiate under the integral sign the normalizing condition

$$\int f_Y(y; \theta) dy = 1$$

so that

$$\int U(\theta; y) f_Y(y; \theta) dy = 0,$$

i.e.,

$$E[U(\theta; Y); \theta] = 0.$$

MLE

Maximum likelihood estimator (MLE): taken here to be $\hat{\theta}$ which solves

$$U(\hat{\theta}; Y) = 0,$$

(or the solution giving largest l if there is more than one)

— a random variable.

We will not discuss (here) situations where the value of θ that maximizes the likelihood is not a solution of the *score equation* as above.

Observed information

Observed information measures curvature (as a function of θ) of the log likelihood:

$$j(\theta) = -\frac{\partial U}{\partial \theta} = -\frac{\partial^2 l}{\partial \theta^2}$$

— the [in general, random] curvature of $l(\theta; Y)$ at θ.

High curvature at $\hat{j} = j(\hat{\theta})$ indicates a well-determined MLE.

Expected information

In most models, $j(\theta)$ is random — a function of Y.

The *expected* information is

$$\begin{aligned} i(\theta) &= E[j(\theta); \theta] \\ &= E\left[-\frac{\partial^2 l}{\partial \theta^2};\ \theta\right] \end{aligned}$$

— a repeated-sampling property of the likelihood for θ; important in asymptotic apprximations.

Expected information is also known as *Fisher information*.

The 'information identity'

We had:

$$\int U(\theta; y) f_Y(y; \theta) dy = 0.$$

Differentiate again under the integral sign:

$$\int \left[\frac{\partial^2 l(\theta; Y)}{\partial \theta^2} + U^2(\theta; Y) \right] f_Y(y; \theta) dy = 0.$$

That is,

$$i(\theta) = \mathrm{var}[U(\theta; Y); \theta].$$

Maximum likelihood can be thought of in various ways as optimal. We mention two here.

The ML 'estimating equation'

$$U(\theta; Y) = 0$$

is an example of an *unbiased estimating equation* (expectations of LHS and RHS are equal).

Subject to some mild limiting conditions, unbiased estimating equations yield consistent estimators.

It can be shown (lecture 7) that the ML equation $U = 0$ is *optimal* among unbiased estimating equations for θ.

Approximate sufficiency of $\{\hat{\theta}, j(\hat{\theta})\}$

Consider the first two terms of a Taylor approximation of $l(\theta)$:

$$l(\theta) \approx l(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2 \hat{j}.$$

Exponentiate to get the approximate likelihood:

$$L(\theta) \approx m(y) \exp[-\frac{1}{2}(\theta - \hat{\theta})^2 \hat{j}],$$

where $m(y) = \exp[l(\hat{\theta})]$.

Interpretation: the pair $(\hat{\theta}, \hat{j})$ is an approximately sufficient statistic for θ.

Re-parameterization

Suppose we change from θ to $\phi(\theta)$ (a smooth 1-1 transformation). This is just a change of the model's coordinate system.

Then:

- $\hat{\phi} = \phi(\hat{\theta})$ — the MLE is unaffected;
- $U^{\Phi}\{\phi(\theta); Y\} = U^{\Theta}(\theta; Y)\frac{d\theta}{d\phi}$ (by the chain rule);
- $i^{\Phi}\{\phi(\theta)\} = i^{\Theta}(\theta)\left(\frac{d\theta}{d\phi}\right)^2$ [since $i = \mathrm{var}(U)$]

The units of information change with the units of the parameter.

Large-sample approximations

It can be shown that (a suitably re-scaled version of) the MLE converges in distribution to a normal distribution.

For this we need some conditions:

- 'regularity' as before (ability to differentiate under the $\int$ sign);
- for some (notional or actual) measure n of the amount of data,
  - $i(\theta)/n \to \bar{i}_\infty$, say, a nonzero limit as $n \to \infty$;
  - $U(\theta)/\sqrt{n}$ converges in distribution to $N(0, \bar{i}_\infty)$.

Asymptotic distribution of $\hat{\theta}$

$$\sqrt{n}(\hat{\theta} - \theta) \to N[0,\ \{\bar{i}_\infty(\theta)\}^{-1}]$$

Sketch proof:

Taylor-expand $U(t; Y)$ around the true parameter value θ:

$$U(t; Y) = U(\theta; Y) - (t - \theta)j(\theta; Y) + \ldots$$

and evaluate at $t = \hat{\theta}$:

$$0 = U(\theta; Y) - (\hat{\theta} - \theta)j(\theta; Y) + \ldots$$

Now ignore the remainder term, re-arrange and multiply by $\sqrt{n}$:

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\frac{U(\theta; Y)}{j(\theta; Y)} = \frac{\frac{1}{\sqrt{n}}U(\theta; Y)}{\frac{1}{n}j(\theta; Y)}.$$

The result follows from the assumptions made, and the fact [based on a weak continuity assumption about $i(\theta)$] that $n^{-1}j(\theta)$ converges in probability to $\bar{i}_\infty$.

$$\sqrt{n}(\hat{\theta} - \theta) \to N[0,\ \{\bar{i}_\infty(\theta)\}^{-1}]$$

So the MLE, $\hat{\theta}$, is distributed approximately as

$$\hat{\theta} \sim N[\theta, i^{-1}(\theta)].$$

Hence approximate pivots:

$$\frac{\hat{\theta} - \theta}{\sqrt{i^{-1}(\theta)}} \quad \text{or} \quad \frac{\hat{\theta} - \theta}{\sqrt{\hat{j}^{-1}}}$$

and approximate interval estimates, e.g., based on $\hat{j}$:

$$\hat{\theta} \pm c\sqrt{\hat{j}^{-1}},$$

with c from the $N(0, 1)$ table.

Three asymptotically equivalent test statistics

Think of testing null hypothesis $H_0: \ \theta = \theta_0$.

Then three possibilities, all having approximately the χ^2_1 distribution under H_0, are:

$$W_E = (\hat{\theta} - \theta_0)i(\theta_0)(\hat{\theta} - \theta_0)$$

$$W_U = U(\theta_0; Y)i^{-1}(\theta_0)U(\theta_0; Y)$$

$$W_L = 2[l(\hat{\theta}) - l(\theta_0)]$$

(the last from a quadratic Taylor approximation to l).

These typically give slightly different results (and W_E depends on the parameterization).

Asymptotic normality of Bayesian posterior distribution

Provided the prior is 'well behaved', the posterior is approximately

$$N(\hat{\theta},\ \hat{j}^{-1}).$$

Multidimensional parameter θ

All of the above results extend straightforwardly. Score is a *vector*, and information is a *matrix*.

Write

$$U(\theta; Y) = \nabla l(\theta; Y).$$

Then

$$E(U) = 0$$

$$\mathrm{cov}(U) = E(-\nabla\nabla^T l) = i(\theta).$$

The extension of the asymptotic normality argument yields

- a *multivariate* normal approximation for $\hat{\theta}$, with variance-covariance matrix $i^{-1}(\theta)$
- test statistics which straightforwardly extend W_E, W_U and W_L.

The information matrix transforms between parameterizations as

$$i^{\Phi}(\phi) = \left(\frac{\partial\theta}{\partial\phi}\right)^T i^{\Theta}(\theta) \left(\frac{\partial\theta}{\partial\phi}\right)$$

and its inverse transforms as

$$\left[i^{\Phi}(\phi)\right]^{-1} = \left(\frac{\partial\phi}{\partial\theta}\right)^T \left[i^{\Theta}(\theta)\right]^{-1} \left(\frac{\partial\phi}{\partial\theta}\right).$$

Nuisance parameters

Suppose $\theta = (\psi, \lambda)$, with ψ of interest.

Then partition vector U into (U_ψ, U_λ), and information matrix (and its inverse) correspondingly:

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}$$

$$i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix}$$

(and similarly for observed information j)

Large-sample results

Simplest route to inference on ψ: approximate normality,

$$\hat{\psi} \sim N(\psi,\ i^{\psi\psi})$$

— from which comes the quadratic test statistic

$$W_E = (\hat{\psi} - \psi_0)^T \left(i^{\psi\psi}\right)^{-1} (\hat{\psi} - \psi_0)$$

[or perhaps use $\left(j^{\psi\psi}\right)^{-1}$ in place of $\left(i^{\psi\psi}\right)^{-1}$].

Corresponding extensions also of W_U and W_L — the latter based on the notion of *profile likelihood*.

Profile likelihood

Define, for any fixed value of ψ, the MLE $\hat{\lambda}_\psi$ for λ.

Then the *profile log likelihood* for ψ is defined as

$$l_P(\psi) = l(\psi, \hat{\lambda}_\psi)$$

— a function of ψ alone.

Clearly $\hat{\psi}$ maximizes $l_P(\psi)$.

The extension of W_L for testing $\psi = \psi_0$ is then

$$W_L = 2\left[l_P(\hat{\psi}) - l_P(\psi_0)\right]$$

— which can be shown to have asymptotically the χ^2 distribution with d_ψ degrees of freedom under the null hypothesis.

Hence also *confidence sets* based on the profile (log) likelihood.

Orthogonal parameterization

Take ψ as given — represents the question(s) of interest.

Can choose λ in different ways to 'fill out' the model. Some ways will be better than others, especially in terms of

- stability of estimates under change of assumptions (about λ)
- stability of numerical optimization.

Often useful to arrange that ψ and λ are *orthogonal*, meaning that $i_{\psi\lambda} = 0$ (locally or, ideally, globally; approximately or, ideally, exactly).

In general this involves the solution of differential equations.

In a full EF, a 'mixed' parameterization is always orthogonal (exactly, globally).

Information in a full EF

Information on the canonical parameters does not depend on Y:

$$i(\phi) = j(\phi) = \nabla\nabla^T k(\phi).$$

So in a full EF model it does not matter whether we use *observed* or *expected* information for inference on ϕ: the answer is the same.

Full EF: Orthogonality of mixed parameterization

If $\phi = (\phi_1, \phi_2)$ and the parameter (possibly vector) of interest is $\psi = \phi_1$, then choosing

$$\lambda = \eta_2 = E[s_2(Y)]$$

makes the interest and nuisance parameters (ϕ_1, η_2) orthogonal.

This follows straight from the transformation rule, for re-parameterization $(\phi_1, \phi_2) \to (\phi_1, \eta_2)$.

Example: The model $Y \sim N(\mu, \sigma^2)$ is a full 2-parameter EF, with $\phi_1 = 1/(2\sigma^2), \phi_2 = -\mu/\sigma^2$ and $(s_1, s_2) = (y^2, y)$. Hence $\mu = E[s_2(Y)]$ is orthogonal to ϕ_1 (and thus orthogonal to σ^2).

Exercise

Let $Y_1, \ldots, Y_n$ have independent Poisson distributions with mean μ. Obtain the maximum likelihood estimate of μ and its variance

(a) from first principles
(b) by the general results of asymptotic theory.

Suppose now that it is observed only whether each observation is zero or non-zero.

- What now are the maximum likelihood estimate of μ and its asymptotic variance?
- At what value of μ is the ratio of the latter to the former variance minimized?
- In what practical context might these results be relevant?

STATISTICAL INFERENCE
Lecture 5
Skeleton notes

1 Asymptotic Bayesian estimation

For Bayesian estimation with a single parameter and a relatively flat prior series expansions show how the f-pivot $(\hat{\theta} - \theta)\sqrt{\hat{j}}$ is approximately also a b-pivot and that departures from the standard normal distribution depend on the asymmetry of the log likelihood at the maximum and the rate of change of the log prior density at the maximum.

Bayesian testing requires more delicate analysis. A key issue is how to specify the dependence, if any, on n of the conditional prior density of θ when the null hypothesis is false.

2 Comparison of test procedures based on log likelihood

There are a considerable number of procedures equivalent to the first order of asymptotic theory, i.e., procedures for which the standardized test statistics agree. For a scalar parameter problem the procedures (all of which appear in the standard software packages) are based

- directly on the log likelihood (Wilks)
- on the gradient of the log likelihood at a notional null point, the score statistic (Rao)
- on the maximum likelihood estimate (Wald)

The last is not exactly invariant under nonlinear transformations of the parameter but is very convenient for data summarization. They would be numerically equal if the log likelihood were quadratic at the maximum. The second does not require fitting a full model and so is especially useful for testing the adequacy of relatively complex models.

The first has the major advantage of retaining at least qualitative reasonableness for likelihood functions of non-standard shape.

3 Jeffreys prior

The notion of a flat and in general improper prior has a long history and some intuitive appeal. It is, however, not invariant under transformation of the parameter, for example from θ to e^θ. The flat priors with most obvious appeal refer to location parameters, so that one resolution of the difficulty is in effect to transform the parameter to approximately location form, take a uniform prior for it and back-transform. This leads to the Jeffreys invariant prior.

Suppose that θ is one-dimensional with expected information $i_\Theta(\theta)$, where the notation emphasizes the parameter under study. Consider a transformation to a new parameter $\phi = \phi(\theta)$. The expected information for ϕ is

$$i_\Phi(\phi) = i_\Theta(\theta)/(d\phi/d\theta)^2.$$

The parameter ϕ has constant information and hence behaves like a location parameter if for some constant c

$$d\phi/d\theta = c\sqrt{i^{\Theta}(\theta)},$$

that is

$$\phi = c\int^{\theta} \sqrt{i^{\Theta}(\kappa)}d\kappa.$$

If now we formally define a flat prior to Φ the prior for Θ is proportional to $d\phi/d\theta$, thus resolving some of the arbitrariness of the notion of a flat prior.

In simple cases this choice achieves second-order matching of frequentist and Bayesian analyses.

For multidimensional problems the Jeffreys prior is proportional to $\{\det(i^{\Theta}(\theta)\}^{1/2}$ but in general it has no obvious optimum properties.

STATISTICAL INFERENCE
Lecture 6
Skeleton notes

1 Outline

Asymptotic theory, Bayesian and frequentist, provides a systematic basis for a wide range of important statistical techniques. There are, however, a number of situations where standard arguments fail and careful analysis is needed. To some extent there are parallel Bayesian considerations. The situations include

- large number of nuisance parameters
- irregular log likelihood
- maximum approached at infinity
- nuisance parameters ill-defined at null point

2 Large number of nuisance parameters

Sometimes called Neyman-Scott problem.
Simplest example is the normal-theory linear model
Methods of resolution

- simplify
- empirical Bayes
- modification of likelihood

For standard normal theory model with $E(Y) = X\beta$ the log likelihood is

$$-n \log \sigma - (y - X\beta)^T (y - X\beta)/(2\sigma^2)$$

which with the least squares estimate defined by $\hat{\beta} = (X^T X)^{-1} X^T y$ can be written

$$-n \log \sigma - \{(y - X\hat{\beta})^T (y - X\hat{\beta}) + (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta)\}/(2\sigma^2).$$

Properties of maximum likelihood estimate of σ^2.
Resolution by factorization of likelihood.

May be possible to apply transformation to new variables V, W such that likelihood is

$$f_V(v; \psi) f_{W|V}(w, v; \psi, \lambda)$$

so that all or nearly all the information about ψ is in the first term. Then use marginal likelihood of V. Alternatively the dependence might be

$$f_V(v; \psi, \lambda) f_{W|V}(w, v; \psi)$$

in which case use the conditional likelihood for inference about ψ.
Application in present example. For inference about σ^2 use marginal likelihood of residual sum of squares.
When would this be inappropriate in both Bayesian and frequentist approaches.

3 Irregular problems

Log likelihood may not be of standard form.

Already discussed multiple maxima.

May be failure of Fisher′s identity and in more extreme form the log likelihood may be discontinuous at maximum.

Simple example.

$Y_1, \ldots, Y_n$ independent and identically distributed in rectangular distribution over $(\theta, 1)$. Likelihood is $1/(1-\theta)^n$ provided $\theta < \min(y_k) = y_{(1)}$ and $\max(y_k) < 1$. Minimal sufficient statistic is $y_{(1)}$. This is within $O_p(1/n)$ of θ.

A more interesting example is that of i.i.d. values from a distribution with, say, a lower terminal, for example

$$\rho \exp\{-\rho(y-\gamma)\}$$

for $y > \gamma$ and zero otherwise.

Similar behaviour. More complicated situations.

Another possibility is that supremum is approached at infinity.

Complete separation in logistic regression

$$\text{pr}(Y_k = 1) = \frac{\exp(\alpha + \beta x_k)}{1 + \exp(\alpha + \beta x_k)}.$$

4 Nuisance parameters ill-defined at null

Simple example

Suppose density is

$$\theta \sigma_1^{-1} \phi\{(y-\mu_1)/\sigma_1\} + (1-\theta)\sigma_2^{-1}\phi\{(y-\mu_2)/\sigma_2\}.$$

Null hypothesis: two components the same.

5 Generalized method of moments

Sometimes may be necessary or helpful not to use likelihood and argue more informally. If the parameter is d-dimensional find d interesting statistics whose expectation can be evaluated under the model. Equate statistics to their expectations and solve.
Generalizations

6 Modified likelihoods

Both Bayesian and frequentist discussions start in principle from the likelihood. There are a number of reasons why modifications of the likelihood may be desirable, for example to produce good frequentist properties or to avoid the need to specify prior distributions over largely unimportant features of the data. Such methods include

- marginal likelihood
- conditional likelihood
- partial likelihood
- pseudo-likelihood
- quasi-likelihood
- empirical likelihood

EXERCISE

Let $Y_1, \ldots, Y_n$ be independently binomially distributed each corresponding to ν trials with probability of success θ. Both ν and θ are unknown. Construct simple (inefficient) estimates of the parameters. When would you expect the maximum likelihood estimate of ν to be at infinity? Set up a Bayesian formulation.

HINT: For the simple estimates, think of two mathematical properties specifying aspects of the binomial distribution, equate these to the corresponding features of the data and solve for an estimate of ν. Are there circumstances in which the estimate is infinite or undefined? Why is this? Suggest a combination of parameter values for which such anomalies are quite likely and simulate say 10 realizations and look at the corresponding likelihoods. When interesting parameter combinations have been found make a more detailed study.

STATISTICAL INFERENCE
Lecture 7
skeleton notes

Non-likelihood inference

Sometimes inference based on likelihood is not possible (e.g., for computational reasons, or because a full probability model cannot be specified).

Sometimes inference based on likelihood may be regarded as not desirable (e.g., worries about impact of failure of tentative 'secondary' assumptions).

Various non-likelihood approaches, including

- 'pseudo likelihoods' — typically designed either for computational simplicity or robustness to failure of (some) assumptions
- 'estimating equations' approaches (includes 'quasi likelihood')

Estimating equations

Consider scalar θ.

Define estimator θ^* as solution to

$$g(\theta^*; Y) = 0$$

— an *estimating equation*, with the 'estimating function' g chosen to that the equation is *unbiased*:

$$E[g(\theta; Y); \theta] = 0$$

for all possible values of θ. (cf. score equation for MLE)

Unbiasedness of the estimating equation results (subject to limiting conditions) in a consistent estimator θ^*.

Examples

Two extremes:

1. Model is fully parametric, $Y \sim f_Y(y;\theta)$. Then the choice $g(\theta;Y) = U(\theta;Y)$ results in an unbiased estimating equation. There may be many others (e.g., based on moments).

2. Model is 'semi-parametric' perhaps specified in terms of some moments. For example, the specification

$$E(Y) = m(\theta)$$

for some given function m may be all that is available, or all that is regarded as reliable: in particular, the full distribution of Y is not determined by θ.

In this case, with Y a scalar rv, the equation

$$g(\theta;Y) = Y - m(\theta) = 0$$

is (essentially) the only unbiased estimating equation available.

Properties

Assume 'standard' limiting conditions. (as for MLE)

Then a similar asymptotic argument to the one used for the MLE yields the large-sample normal approximation

$$\theta^* \sim N\left(\theta,\ \frac{E(g^2)}{[E(g')]^2}\right).$$

Note that the asymptotic variance is invariant to trivial scaling $g(\theta;Y) \to ag(\theta;Y)$ for constant a — as it should be, since θ^* is invariant.

Lower bound on achievable variance

(Godambe, 1960)

For unbiased estimating equation $g = 0$,

$$\frac{E(g^2)}{[E(g')]^2} \geq \frac{1}{E(U^2)} = i^{-1}(\theta),$$

where $U = \partial \log f / \partial\theta$.

Equality if $g = U$.

This comes from the Cauchy-Schwarz inequality; it generalizes the Cramér-Rao lower bound for the variance of an unbiased estimator.

A simple illustration

Suppose that counts Y_i ($i = 1, \dots, n$) are made in time intervals t_i.

Suppose it is suspected that the counts are over-dispersed relative to the Poisson distribution. The actual distribution is not known, but it is thought that roughly $\text{var}(Y_i) = \phi E(Y_i)$ (with $\phi > 1$).

Semi-parametric model:

1. $E(Y_i) = t_i r(x_i; \theta) = \mu_i$
2. $\text{var}(Y_i) = \phi \mu_i$.

The first assumption here defines the parameter of interest: θ determines the rate (r) of occurrence at all covariate settings x_i.

The second assumption is more 'tentative'.

Hence restrict attention to estimating equations unbiased under only assumption 1: don't require assumption 2 for unbiasedness, in case it is false.

Use assumption 2 to determine an *optimal* choice of g, among all those such that $g = 0$ is unbiased under assumption 1.

Consider now the simplest case: $r(x_i, \theta) = \theta$ (constant rate).

The possible unbiased (under 1.) estimating equations are then

$$g(\theta; Y) = \sum_1^n a_i (Y_i - t_i \theta)$$

for some choice of constants $a_1, \dots, a_n$.

Using both assumptions 1 and 2 we have that

$$\frac{E(g^2)}{[E(g')]^2} = \frac{\sum a_i^2 \phi t_i \theta}{(\sum a_i t_i)^2}$$

— which is minimized when $a_i = \text{constant}$.

The resulting estimator is $\theta^* = \sum Y_i / \sum t_i$ (total count / total exposure)

— which is 'quasi Poisson' in the sense that it is the same as if we had assumed the counts to be Poisson-distributed and used MLE. (But standard error would be inflated by an estimate of $\sqrt{\phi}$.)

— a specific (simple) instance of the method of 'quasi likelihood'.

Some generalizations:

- vector parameter
- working variance $\longrightarrow$ working variance/correlation structure: quasi-likelihood $\longrightarrow$ 'generalized estimating equations'
- estimating equations designed specifically for outlier robustness

etc., etc.

APTS module *Statistical Inference*
D. R. Cox and D. Firth
December 2007

Some suggested reading/reference material

Books

Cox, D. R. (2006). *Principles of Statistical Inference.* CUP.
    Closest book to the APTS lectures.

Cox, D. R and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall.
    An older and more detailed account of similar material.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference.* CUP.
    A broad, concise introduction, including some decision analysis.

Lehmann, E. L. and Romano, J. P. (2004, 3rd ed). *Testing Statistical Hypotheses.* Springer.
Lehmann, E. L. and Casella, G. C. (2001, 2nd ed). *Theory of Point Estimation.* Springer.
    Much detailed mathematical material on estimation and testing.

Jeffreys, H. (1961, 3rd ed.). *Theory of Probability.* OUP.
    A detailed development of objective Bayesian theory.

Savage, L. J. (1954). *The Foundations of Statistics.* Wiley.
    Pioneering account of the personalistic Bayesian view.

O'Hagan, A. and Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics: Bayesian Inference.* Arnold.
    Thorough treatment of current Bayesian approaches.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* Wiley.
    A good discussion of the principles and methods of decision analysis.

Edwards, A. W. F. (1972). *Likelihood.* CUP.
    Statistical theory based solely on likelihood.

McCullagh, P. and Nelder, J. A. (1989, 2nd ed.). *Generalized Linear Models.* Chapman and Hall.
    An authoritative account of the generalized linear model (links with the APTS module *Statistical Modelling*).

Papers

Fisher, R. A. (1950). The significance of deviations from expectation in a Poisson series. *Biometrics* **6**, 17-24.
    Conditional test of lack of fit: as in lecture 2.

Cox, D. R (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357-372.
    A discussion of conditioning and the relations between various approaches.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Stat. Soc.* B **34**, 1-41.
    Influential paper on hierarchical Bayesian analysis.

Joe, H. and Reid, N. (1985). Estimating the number of faults in a system. *J. Amer. Stat. Assoc.* **80**, 222-226.
    Related to assessment exercise B(2) on inference for the binomial with unknown index.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Stat. Soc.* B **49**, 1-39.
    Orthogonality (lecture 4).

McCullagh, P. (1991). Quasi-likelihood and estimating functions. Pages 265-286 in: Hinkley DV, Reid N and Snell EJ (eds.), *Statistical Theory and Modelling, in Honour of Sir David Cox FRS.* Chapman and Hall.
    Review of the theory of estimating equations.